Paper / Subject Code: 42476 / BIG DATA ANALYTICS (DLOC - III)

Duration: 3hrs **[Max Marks:80]**

N.B. : (1) Question No 1 is Compulsory.
(2) Attempt any three questions out of the remaining five.
(3) All questions carry equal marks.
(4) Assume suitable data, if required and state it clearly.

1 Attempt any FOUR **[20]**

A What is Big Data? What is Hadoop? How are Big Data and Hadoop linked?

B Write the step of Grivan-Newman algorithm. Explain clustering of Social Network Graph using GN algorithm with example.

C What is MapReduce ? Explain How Map and Reduce Work?

D Explain PCY algorithm with suitable examples.

E Explain NoSQL data Architecture patterns.

F Explain Recommendation system & its various types with example.

2 a Describe the structure of HDFS in a Hadoop Ecosystem using a diagram **[10]**

b What is NOSQL? What are the business drivers for NoSQL? Discuss any two architectural patterns of NoSQL. **[10]**

3 a Explain Page Rank with Example. Can a Website's Page rank Ever Increase? What are its chances of Decreasing? **[10]**

b Evaluate PCY algorithm on the following transaction to find the candidate sets (frequent sets). **[10]**

Given data: Threshold value or minimization value = 3

Hush function = (i * j) mod 10.

T1 = {1, 2, 3}	T2 = {2, 3, 4}	T3 = {3, 4, 5}
T4 = {4, 5, 6}	T5 = {1, 3, 5}	T6 = {2, 4, 6}
T7 = {1, 3, 4}	T8 = {2, 4, 5}	T9 = {3, 4, 6}
T10 = {1, 2, 4}	T11 = {2, 3, 5}	T12= {3, 4, 6}

4 a Explain the Role and effect of damping Factor(teleportation) in page rank computation **[10]**

b Calculate the Cosine distance measure for given vectors **[10]**

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

5 a Explain Clearly with diagram how the PCY algorithm helps to perform frequent itemset mining for large datasets **[10]**

b Give the formal definition of Nearest Neighbor problem,Show how finding plagiarism in a document is nearest Neighbour Problem. What similarity measure can be used **[10]**

6 a Given a Dim Dataset (1,5,8,10,2} Use the agglomerative clustering algorithm with Euclidean distance to establish hierarchical grouping relationship. Draw the dendrogram. **[10]**

b Write a note on (Any Two) **[10]**

i) HITS

ii) Distance measurement for Big data

iii) Multistage Frequent Itemset Mining Algorithm
